AoI - Action on Invasives Data Policy & Strategy

Policy

We recognise that there are a range of potential barriers to access to scientific knowledge in agriculture and the environment, especially in areas where such knowledge can have most benefit: developing nations.

The Action on Invasives (AoI) programme supports the general principles laid out in CABI's published open access policy[1]. **AoI supports 100% Open Access publishing of all key scientific and development-relevant publications that arise from this programme[2]**. Where possible **Gold Open Access is preferred where outputs are in peer-reviewed publications**.

However, we also contend that It is not enough just to mandate for Open Access to scientific research in the conventional sense (that is removal of pay barriers to peer-reviewed science journal articles), rather CABI recognises that all research outputs – data, information, and knowledge should be created, assembled, handled and communicated in ways that ensure that they will be made available to all that need them. Barriers to knowledge access can be many and varied and include language, internet connectivity, poor website design, poor indexing of data assets (for people and machines), poor search engine discoverability, and poorly articulated research communication, for gender or youth. We will design AoI programmatic interventions with such barriers in mind. Where we support development of systems that would normally be subscription products (such as Pest Risk Analysis through the Crop Protection Compendium) we do so only where there is an access model to support all those who reasonably might be considered unable to pay for access. Data outputs from such systems should be easily reusable and not restricted by licence.

We aim to consider how data is made available in this broader context, and to that extent we recommend that **all AoI programme data output should aim to be 'FAIR' - that is Findable, Accessible, Interoperable and Reusable'[3].** In line with these principles we will ensure that **all key data assets arising from the programme's activity should be published,** and **all Open Access publications should be supported by published data sets.**

Programmatic policy on data management and publishing support agreed targets for the programme, most notably contained in its Logframe. Key indicators there explicitly require that we agree on processes for exchange of content and data (Activity 4.4), develop online and offline systems for managing two way flow of Information and Data (Activity 4.5). Discussion of these priorities led to convening of a Knowledge & Data-focused Stakeholder Workshop in

[1] https://www.cabi.org/Uploads/CABI/about-us/4.8.5-other-business-policies-and-strategies/Data%20management.pdf

[2] The CABI Science Strategy says that we will publish only in open access spaces by 2019 https://www.cabi.org/Uploads/CABI/about-us/4.8.5-other-business-policies-and-strategies/Science%20strategy.pdf

[3] https://www.force11.org/group/fairgroup/fairprinciples

November 2017 which suggested an approach to Data Management Planning which informs the Strategy laid out below.

We accept a fairly broad definition of what an invasives dataset is; that is one which is likely to be of interest to invasive species experts. That may include datasets more clearly core to those aims such as data on introductions, species spread, and pathways, but could be broader.

Strategy

The scope of the strategy is framed by two key questions:

1. what data is the programme is creating and who we want to share it with and for what purpose?
2. what other data is relevant to invasive species and needed by those who can support improved science, decision making in this space (modellers, etc)

Background

This Strategy is intended to be a practical guidance document; a series of clear recommendations and plans, focussing on a mix of quick wins and medium to long term recommendations. The final version should be prioritised (& lightly costed where possible). We recognise that any recommendations for implementation beyond what can be achieved in the middle of 2019 may not be funded so we aim to think modularly, driving improvements to how we manage and share data in real use cases and existing interventions as defined in the programme - starting small but thinking big.

Any platforms developed to support improved data management, publishing and use should where possible have cross programme applicability, and ideally act to improve the data processes and infrastructure for CABI and its partners more generally beyond the requirements of the programme.

Approach to Improving AoI Data Management

The AoI programme seeks to promote:

- **Data storage and preservation for future use**
    - We seek to set up suitable repositories for AoI data outputs and to promote the generation of new shared spaces around the community of Invasive Species experts that we seek to serve. In addition, we will modify existing platforms to support increased availability of data.
    - **We will build a CKAN instance interfacing with Open Data Kit** for AoI and its intended user base, seeking to produce a searchable catalogue (inventory) of key assets; publish data sets (or at least metadata records for data sets) from

    CABI’s (currently internal) dataset inventory, plus external datasets or metadata records and links of relevance to the sector (supported by Compendia team);
  - and develop a **new Distribution Database** which can drive improved publishing of standardised, increasingly granular, and well indexed distribution data to serve CABI’s data-driven products including the Open Access **Invasive Species Compendium**.

- **Data use**
  - We will build a **user community** around our new CKAN repository and develop it iteratively responding to their needs. We have identified a community of modellers around *Parthenium* (in Pakistan and internationally) who have a requirement for a central open repository to share survey and distribution data. This will be a good area to focus on initially. Our own M&E team would be another group. Needs of other user communities (sometimes around key species, sometimes in named geographies can be surveyed and interrogated on their needs as budgets and resources dictate; this is a fundamentally scalable activity).
  - We will **survey understanding of opportunities and challenges** around data sharing only in selected communities of invasives experts and collect inventories of locally relevant datasets such experts may be aware of. Trial in Parthenium-interest community in Pakistan and review results before considering scaling (to Kenya, Zambia or Ghana). Continue to piggy back activities where possible on Plantwise assessments of plant health system and clinic data sharing challenges.
  - We will **identify clusters of datasets that are likely to be of interest** to scientists and modellers of invasive species and focus on publishing those on CKAN with **specific analyses** in mind, and **conducting analyses for publication**.
  - We will add commenting functions on CKAN and encourage its use
  - We will generate simple guidelines for data publication generally and CKAN use specifically. We will support CABI and community scientists to understand how to use CKAN effectively, how to use it to publish internally (in CABI or defined communities), or publicly, and how to recognise and handle sensitive data. Where CKAN is not preferred for whatever reason (some other repositories may be better for certain datasets) scientists should be supported to openly publish datasets there instead and a linking metadata record should be held on CKAN.
  - Where CKAN records do not resolve to a data set directly there should be a clear route identified as to where the dataset may be obtained (active email contact, etc.). We should seek to monitor responsiveness to requests for data.
  - To promote data understanding **we will add tools to CKAN to make visualization of data easier**, especially visualization of survey data
  - We will **launch CKAN to our partners through our media channels** in coordination with CABI marketing and comms teams. We will focus on telling a story around the data, possibly with a focus on Parthenium in Pakistan.

    - All **Evidence Notes produced by the programme will publish key datasets** to support the narrative in CKAN
    - **All peer reviewed publications funded under AoI will be Open Access and will link to all relevant supporting / underlying datasets** on a reputable open repository (e.g. GBIF or Dataverse) or our own CKAN repository. Where a repository other than our CKAN is used a metadata entry on our own CKAN instance will link to the dataset at source.
    - CABI scientists publishing or supporting inventorisation of data sets in support of AoI will be supported to do so with provision of data analyst time, expertise and where possible mentoring to improve and embed good data management practice in our invasives scientists (named individuals to be identified).

- **Good quality metadata to improve data discoverability**
    - All data on our CKAN instance will be published to an agreed metadata structure model which will identify owners, licencing terms and recommended forms of **data citation.** Use of standardised vocabularies, schemas and ontologies will promote interoperability.
    - We will use other well indexed datasets to improve the accuracy of the Compendium Horizon Scanning Tool.

- **FAIR and Open Data**
    - We seek to make AoI data assets increasingly FAIR, that is Findable, Accessible, Interoperable and Reusable. What this means in practice is covered in most part elsewhere in this strategy, but we acknowledge also that we can learn from and adopt approaches for adopting FAIR from other projects that CABI is active in as opportunity presents itself.[4]
    - We agree to publish or make available for download data sets in non-proprietary formats where possible.
    - We endorse machine readable data solutions and seek to develop APIs on AoI-endorsed platforms where appropriate to do so.
    - All published AoI data assets will carry a clear licence with a clear preference for open licences such as those available under Creative Commons (CC).
    - We will work with the product owners of the Invasive Species Compendium and Crop Protection Compendium to agree on the most appropriate Creative Commons licence to adopt for collaborative outputs (e.g. CC-BY for PRA reports). This may necessitate removal of ND and or NC from standard Compendia licences where appropriate.

- **Interoperability**
    - Develop clear lists of data types that our platforms will handle.
    - Use of standardised shared languages, vocabularies, schemas and ontologies to promote interoperability in the broader community, not just inside CABI.

[4] Smith F, Dodds L, Day C *et al.* Creating FAIR and open data ecosystems for agricultural programmes [version 1; not peer reviewed]. *Gates Open Res* 2018,**2**:42 (document) (https://doi.org/10.21955/gatesopenres.1114883.1)

    - Use standardised descriptive frameworks that are available under an open licence and which follow FAIR principles (such as GACS).
- User data, ethics, confidentiality, anonymization
    - We agree to responsible data use[5]
    - Consent to be surveyed using e.g ODK should be informed
    - Implications of use of the PRA tool - that CABI will use data in certain ways should be made clear to users
    - Consent to share PRA data outputs should be sought; users should be able to opt out and encouraged to opt in (and the advantages of doing so explained).
    - We collect data in such a way as to appropriately sample and survey the perspectives and lives of women. All survey data should be gender disaggregated. Any analyses should consider whether gender sampling has been appropriate and seek to report on and correct for where it has not been.
    - We are aware that invasive species datasets will often include distribution records for species in a country where it has not previously been reported. We are committed to working alongside National Plant Protection Organisations (NPPOs) and, where possible, facilitating the IPPC contracting parties' obligation to report new pest records. In order to support this, we will engage with NPPOs and other key stakeholders to better understand how they would want to view the data and provide feedback.
- Data metrics for monitoring and evaluation of performance
    - We will add usage metrics to CKAN to track reads and downloads before its release and promotion.
    - CKAN links will be made available from species data portals (FAW and others) and click throughs from these entry points will be monitored.
    - Usage data will be used to inform building of better platforms.
    - Usage data on all platforms and tools may be used in reports and peer-reviewed publications to promote to others how our AoI investments have led to access (& use) of materials (PRA, ISC, HST, etc).

Notes on implementation of Strategy

CKAN

"*CABI will invest in capturing its research data outputs in an accessible and usable way, so that they can be reused and shared as open data"*[6]

CABI's CKAN data repository will be developed to serve AoI's and CABI's growing data collection, discovery and publishing needs.

- AoI activities will continue to use CKAN and ODK for collecting data in the field.

[5] For example, frameworks such as https://theodi.org/article/data-ethics-canvas/ should be consulted when trying to render and publish anonymised data from surveys, etc.

[6] CABI Science Strategy

- More features will be added to the CKAN site in 2019 to allow users to visualise data in the repository as it is being collected. Users will be able to build custom graphs and charts to monitor flows of incoming data and view some basic analysis
- CKAN data collection tools will be promoted to other areas of CABI who may benefit from the new capabilities. Probable areas include the M&E team and field scientists
- The AoI CKAN data repository will continue to be populated with metadata on key CABI invasive species datasets and where suitable, the datasets can be published openly and made available to download.
- The AoI CKAN data repository will continue to be populated with metadata on key external sources of invasive species datasets
- CABI will publish and link to the CKAN repository through other channels e.g. marketing materials, blogs, links from other products such as the ISC and species portals.
- The AoI team will establish ‘community data repositories’ on CKAN. A demand has been identified for collections of distribution data on key species such as Parthenium and Fall Armyworm to meet the needs of modellers.

Distribution Database

- Phase 1 of the Distribution Database (DDB) has produced a CABI-wide data model and a single structure for handling and storing CABI’s distribution data. A user interface has been produced to allow CABI staff to view, edit and manage distribution records. This work has included linking to and aligning with other key CABI data sources such as CAB Thesaurus and CAB Abstracts.
- Phase 2 is identified as a strategic investment for CABI and will be co-funded using also CABI core funding and Compendia consortium funds.
- Phase 2 of the DDB project will populate the database with CABI’s data assets including the migration of Compendia and Plantwise distribution records. Then all of CABI’s existing products will be connected to the new database resource. This includes Compendia products (including the ISC), Horizon Scanning Tool, Pest Risk Assessment Tool and the Plantwise Knowledge Bank.
- Further enhancements will be made to provide users with better tools in order to bulk upload datasets and semi-automated management of summary and geographically inferred records.
- At the completion of Phase 2 CABI will have a strong foundation on which to:
    - Import data from external sources
    - Create better maps and data visualisations in products
    - Quickly build distribution and geographic features into products
    - Explore new product opportunities
- Phase 3 will be lightly scoped and costed so that we can implement further improvements should additional funds become available.

HST and PRA tool

The AoI has supported the development of the beta version of the PRA tool, and will support development through 2019 to full launch in Q4. Metrics on PRA usage will be gathered[7]. Data on usage from the PRA will be used in the improvement of programme design and M&E. The export forms from the PRA[8] will be available under a CC-BY license where authors have chosen to share data beyond CABI.

We will review ISC licenses and support development of data sharing (and data sharing agreements) with key collaborating parties as necessary.

Based on user feedback, a demand has been identified for the HST and PRA tool to 'rank' the species returned in list of search results. In conjunction with the SAUKOT PRA project the AoI team will develop a method for prioritizing species based on risk. For this to be fully applied in the HST and PRA tool we will need to seek further funding.

Evidence Notes

Evidence notes will be openly published on the CKAN site alongside supporting datasets.

AoI Peer Reviewed Science

AoI peer reviewed publications will be published on the CKAN site and where possible the supporting datasets will be published alongside the research.

Sustainability

The AoI programme seeks to develop approaches, tools and deliverables which will achieve longevity beyond the programme by virtue of becoming embedded within the ways we work at CABI.

CKAN and the Distribution database are cases in point. Both support AoI but even in their pilot stage are designed to be prototypes for CABI infrastructures for the management of data, especially geospatial data.

Publication of all key deliverables in the open (OA & OD) on robust platforms will ensure that the findings of the research from AoI will be available to CABI and the global invasives community long beyond the funded investment.

Delivery of data-driven tools that are embedded in CABI products (HST and PRA) gives them inherent longevity beyond the end of the programme.

[7] PRA users will be given insight as to how their data will be used to ensure informed consent.

[8] https://www.cabi.org/PRA-Tool/

Documentation of usage of key tools (e.g. species portal, ISC, PRA) and how they have increased as a result of the programme will include collection of quantitative and qualitative insights. Case studies of PRA usage will be used to brief donors on programme impact, and to support marketing and communication of spin offs such as commercial use of the PRA.

Development of a team of data analysts in the course of the project will allow the expertise to be institutionalized, and commercialized as service offerings deployed in follow on initiatives.